Amendments to the Specification:

Please replace the section of the specification from page 163, line 1, to page 169, line 15, with the following redlined section:

In an embodiment using cluster analysis, the transcription levels of a number of transcription control sequences can be monitored while applying various stimuli to biological samples. A table of data containing measurements of the transcription levels of transcription control sequences is used in cluster analysis. In order to obtain a set of basic transcription control sequences containing transcription control sequences which simultaneously vary under various conditions, typically at least two, preferably at least 3, more preferably at least 10, even more preferably more than 50, and most preferably more than 100 stimuli or conditions are used. Cluster analysis is performed for a table of data having m×k dimensions where m is the total number of conditions or stimuli and k is the number of transcription control sequences to be measured.

A number of clustering algorithms are useful for clustering analysis. In clustering algorithms, differences or distances between samples are used to form clusters. In a certain embodiment, a distance used is a Euclidean distance in multi-dimensional space:

$$I(x,y)=\left\{\sum_i (X_i - Y_i)^2\right\}^{1/2} \qquad (1)$$

where (x, y) represents a distance between gene X and gene Y (or any other cellular components X and Y (e.g., transcription control sequences)); X_i and Y_i represent gene expression in response to i stimuli. Euclidean distances may be squared and then multiplied with weighting which are increased with an increase in the distance. Alternatively, a distance reference may be, for example, a distance between transcription control sequences X and Y, or a Manhattan distance represented by:

$$I(x,y)=\sum_i |X_i - Y_i| \qquad (2)$$

where X_i and Y_i represent responses of transcription control sequences or gene expression when i stimuli are applied. Several other definitions of distance include Chebyshev distance, power distance, and mismatch rate. When dimensional data can be categorized without modification, a mismatch rate defined as I(x, y) = (the number of $X_i \neq Y_i$)/i may be used in a method of the present invention. Such a method is particularly useful in terms of cellular responses. Another useful definition of distance is I=1-r where r is a correlation coefficient of response vectors X and Y, e.g., a normalized inner product X·Y/|X||Y|. Specifically, an inner product X·Y is defined by:

$$X \cdot Y = \sum_i X_i \times Y_i \qquad (3).$$

Also, $|X|=(X \cdot X)^{1/2}$ and $|Y|=(Y \cdot Y)^{1/2}$.

Most preferably, a distance reference is suited to a biological problem in order to identify cellular components (e.g., transcription control sequences, etc.) which are simultaneously changed and/or simultaneously regulated. For example, in a particularly preferred embodiment, a distance reference is I=1-r having a correlation coefficient containing a weighted inner product of genes X and Y. Specifically, in such a preferred embodiment, [[~~r~~]]r_n is defined by:

$$r = \frac{\sum_i \frac{X_i Y_i}{\sigma_i^{(X)} \sigma_i^{(Y)}}}{\left[\sum_i \left(\frac{X_i}{\sigma_i^{(X)}} \right)^2 \left(\frac{Y_i}{\sigma_i^{(Y)}} \right)^2 \right]^{1/2}} \qquad (4)$$

where $\sigma_i^{(X)}$ and $\sigma_i^{(Y)}$ represent standard errors in measurement of genes X and Y in experiment i.

The above-described normalized and weighted inner products (correlation coefficients) are constrained between values +1 (two response vectors are completely correlated, i.e., the two vectors are essentially the same) and -1 (two response vectors are not correlated or do not have the same orientation (i.e., opposing orientations)). These correlation coefficients are particularly preferable in an embodiment of the present invention which tries to detect a set or cluster of cellular components (e.g., transcription control sequences, etc.) having the same sign or response.

In another embodiment, it is preferable to identify a set or cluster of cellular components (e.g., transcription control sequences, etc.) which simultaneously regulate the same biological response or pathway or are involved in such regulation, or have similar or non-correlated responses. In such a embodiment, it is preferable to use the absolute value of either the above-described normalized or weighted inner product, i.e., |r| as a correlation coefficient.

In still another embodiment, the relationship between cellular components (e.g., transcription control sequences, etc.), which are simultaneously regulated and/or simultaneously changed, are more complicated, e.g., a number of biological pathways (e.g., signal transduction pathways, etc.) are involved with the same cellular component (e.g., a transcription control sequence, etc.) so that different results may be obtained. In such an embodiment, it is preferable to use a correlation coefficient $r=r^{(\mathrm{change})}$ which can identify cellular components (other transcription control sequences as controls which are not involved in change) which are simultaneously changed and/or simultaneously regulated. A correlation coefficient represented by expression (5) is particularly useful for the above-described embodiment:

$$r = \frac{\sum_i \left| \frac{X_i}{\sigma_i^{(X)}} \right| \left| \frac{Y_i}{\sigma_i^{(Y)}} \right|}{\left[\sum_i \left(\frac{X_i}{\sigma_i^{(X)}} \right)^2 \left(\frac{Y_i}{\sigma_i^{(Y)}} \right)^2 \right]^{1/2}} \qquad (5).$$

Various cluster linkage methods are useful in a method of the present invention.

Examples of such a technique include a simple linkage method, a nearest neighbor method, and the like. In these techniques, a distance between the two closest samples is measured. Alternatively, in a complete linkage method, which may be herein used, a maximum distance between two samples in different clusters is measured. This technique is particularly useful when genes or other cellular components naturally form separate "clumps".

Alternatively, the mean of non-weighted pairs is used to define the mean distance of all sample pairs in two different clusters. This technique is also useful in clustering genes or other cellular components which naturally form separate "clumps". Finally, a weighted pair mean technique is also available. This technique is the same as a non-weighted pair mean technique, except that in the former, the size of each cluster is used as a weight. This technique is particularly useful in an embodiment in which it is suspected that the size of a cluster of transcription control sequences or the like varies considerably (Sneath and Sokal, 1973, "Numerical taxonomy", San Francisco: W.H. Freeman & Co.). Other cluster linkage methods, such as, for example, non-weighted and weighted pair group centroid and Ward's method, are also useful in several embodiments of the present invention. See, for example, Ward, 1963, J. Am. Stat. Assn., 58: 236; and Hartigan, 1975, "Clustering algorithms", New York: Wiley.

In a certain preferred embodiment, cluster analysis can be performed using a well-known hclust technique (e.g., see a well-known procedure in "hclust" available from Program S-Plus, MathSoft, Inc., Cambridge, MA).

According to the present invention, it was found that even if the versatility of stimuli to a clustering set is increased, a state of a cell can be substantially elucidated by analyzing typically at least two, preferably at least 3, profiles using a method of the present invention. Stimulation conditions include treatment with a pharmaceutical agent in different concentrations, different measurement times after treatment, response to genetic mutations in various genes, a combination of treatment of a pharmaceutical agent and mutation, and changes in growth conditions (temperature, density, calcium concentration, etc.).

As used herein, the term "significantly different" in relation to two statistics means that the two statistics are different from each other with a statistical significance. In an embodiment of the present invention, data of a set of experiments assessing the responses of cellular components can be randomized by a Monte Carlo method to define an objective test.

In a certain embodiment, an objective test can be defined by the following technique. p_{ki} represents a response of a component k in experiment i. $\Pi_{(i)}$ represents a random permutation of the indices of experiments. Next, $p_{k\Pi(i)}$ is calculated for a number of different random permutations (about 100 to 1,000). For each branch of the original tree and each permutation:

(1) hierarchical clustering is performed using the same algorithm as that which has been used for the original data which is not permutated (in this case, "hclust"); and

(2) an improvement f in classification in total variance about the center of clusters when transition is made from one cluster to two clusters;

$$f = 1 - \Sigma D_k^{(1)} / \Sigma D_k^{(2)} \qquad (6).$$

where D_k is the square of the distance reference (mean) of component k with respect to the center of a cluster to which component k belongs. Superscript 1 or 2 indicates the center of all branches or the center of the more preferable cluster of the two subclusters. The distance function D used in this clustering technique has a considerable degree of freedom. In these examples, D=1-r, where r is a correlation coefficient of one response with respect to another response of a component appearing in a set of experiments (or of the mean cluster response).